DOCUMENT RESUME

ED 138 272 IR 004 693

AUTHOR Blashfield, Roger
TITLE A Guide to Cluster Analysis Software.
PUB DATE 75
NOTE 10p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Classification; *Cluster Analysis; *Computer Programs; *Programing Languages; Statistical Analysis

ABSTRACT

This paper briefly summarizes the characteristics of 14 of the currently most popular computer pragrams for cluster analysis. Also provided is a subjective ranking of the programs in terms of their versatility and usability. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED138272

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A Guide to Cluster Analysis Software

Roger Blashfield
Department of Psychiatry
University of Florida

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY

Roger K. Blashfield

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

Abstract

The interest in cluster analysis has grown rapidly since the publication by Sokal & Sneath (1963). This article briefly summarizes the characteristics of 14 of the currently most popular computer programs for cluster analysis. Also provided is a subjective ranking of the programs in terms of their versatility and useability.

IR004693

A Guide to Cluster Analysis Software

Roger K. Blashfield
Department of Psychiatry
University of Florida

This paper concerns computer software programs which embody various methods of cluster analysis. These quasi-statistical methods are used to form groups of relatively homogeneous entities and is potentially useful to any science concerned with the creation of classifications. The literature on cluster analysis has expanded greatly since the early 1960's. Much of the interest in this technique was stimulated by the work on numerical taxonomy by the biologists, Sokal and Sneath (1963). In their first book, they provided an outline of a number of cluster analysis methods, a rationale of the relevance of this family of methods for biological classification and a revision to the standard taxonomic theories about classification.

Since the early 1960's there have been over thirty different books and journal articles published which focus primarily on reviewing cluster analysis methods. Neophytes to this literature are recommended to read the introductory reviews by Bailey (1974) and by Everitt (1974). More sophisticated readers will find the reviews of Anderberg (1973), Cormack (1971), Hartigan (1975) and Sneath & Sokal (1973) valuable.

A review of the literature on cluster analysis will support the following four conclusions: (a) there are a large number of clustering methods, and these various methods can and often do generate different solutions to the same data (Everitt, 1974; Sneath & Sokal, 1973): (b) there is no clearly preferrable method of cluster analysis since the literature comparing the various methods is conflicting (Blashfield, 1976; Cunningham & Ogilvie, 1972; Sibson, 1971; Williams, Lance, Dale & Lambert, 1971); (c) there are a large number of sciences interested in cluster analysis and as a result no standard terminology for discussing this family of procedures exists (Aldenderfer & Blashfield, 1976; Anderberg, 1973): and (d) although a number of different theoretical approaches to cluster analysis have been proposed, none have gained ascendency (Bezedek, 1974; Hubert, 1972; Jardine & Sibson, 1971; Ling, 1972; Peay, 1975; Tryon & Bailey, 1970; Wolfe, 1970).

Given these sources of marked variation in the study of cluster analysis, one can expect that there will also be a great deal of variation in the computer software which involve cluster

analysis. Different programs and packages are likely to be created by persons whose perspectives about cluster analysis are different.

The number of cluster analysis programs has grown rapidly. In a survey questionnaire sent to 167 scientists interested in classification (Blashfield, in press), the 53 respondents listed 50 different programs for performing cluster analysis. If this result can be generalized, it would seem that there are as many programs for performing cluster analysis as there are users. From the survey of scientists interested in classification, 14 cluster analysis programs were the most popular.

The 14 programs are discussed in detail below. The reader will note that the programs are classified into three categories: packages which contain a number of clustering and statistical procedures, programs which are designed to perform only one type of clustering method and books that contain a number of FORTRAN subroutines which can be combined to form a package of cluster analysis methods.

I. General Statistical Packages Containing Cluster Analysis Methods

BCTRY -- This package is based upon the extensive theoretical work by R. C. Tryon who published the first book on cluster analysis in 1939. It should be noted that what Tryon meant by "cluster analysis" is what most psychometricians would describe as "multiple group factor analysis." This package is primarily designed to contain a large range of options involving factor analysis. Two clustering methods are contained in BCTRY: a) an iterative partitioning method which emphasizes procedures for merging clusters and b) a clustering routine related to inverse factor analysis.

BMDP -- This general statistical package is a recent version of the popular BIOMED statistical packages. The major innovation in BMDP is a highly readable, free format, mnemonic control language. The BMDP package contains routines for performing simple data description, correlation, factor analysis, cluster analysis, etc. There are three procedures in BMDP for cluster analysis: BMDP1M has three hierarchical agglomerative methods for clustering variables; BMDP2M has one hierarchical agglomerative method for clustering entities; and BMDP3M simultaneously forms clusters of both entities and variables.

CLUSTAN -- This collection of cluster analysis routines was formed in Great Britain by David Wishart. Currently it is the only statistical package which exclusively focuses on cluster

analysis. This package contains hierarchical agglomerative, hierarchical division, iterative partitioning, mode searching and minimum spanning tree methods. CLUSTAN is the most versatile of the programs in terms of hierarchical agglomerative methods. Thirty-eight similarity/dissimilarity measures and eight different agglomerative methods are available.

NT-SYS -- Intended for users in biology, NT-SYS is a package of multivariate statistical techniques including cluster analysis, factor analysis and multidimensional scaling. The cluster analysis routines emphasized in the package are those popularized by Sokal & Sneath (1963). Five different hierarchical agglomerative and 21 similarity/dissimilarity measures are available. This package has excellent options concerning the graphics output for hierarchical solutions.

OSIRIS -- OSIRIS is a general statistical package primarily intended for use in the social sciences. The program contains a number of routines for data description, tabulation and editing. There are two cluster analysis routines in OSIRIS: CLUSTER which contains a clumping routine and HICLSTR which has one hierarchical agglomerative method. Both routines are designed for clustering variables. In addition, there is a third program, AID3, which uses a monothetic hierarchical divisive method to form groups which optimally predict a criterion variable.

II. Programs Specifically Concerned with One Type of Cluster Analysis

BUILDUP -- This program was developed by Lorr & Radhakrishnan (Educational and Psychological Measurement, 1967) to embody a clustering algorithm which has been used to form classifications of psychotic patients. The program requires the user to input a correlation matrix, and the algorithm forms one cluster at a time. The user has options for controlling the homogeneity within a cluster.

CLUS -- This program is based upon an article in Journal of American Statistical Association (1967) by Friedman & Rubin. The program only contains iterative partitioning methods and emphasizes "hill climbing" passes to optimize statistics used in one-way multivariate analysis of variance. CLUS is the most versatile program for iterative partitioning methods. Options are provided for analyzing either pooled covariance structures or performing a partitioning analysis of a similarity matrix. Other major options in CLUS concern the choice of the starting partition, the type of membership reassignment procedure and the statistic to be optimized.

HGROUP -- This program is contained in Veldman's collection of statistical routines. (Veldman, D. J., FORTRAN Programming for the Behavioral Sciences. Holt, Rinehart & Winston, 1967). The program performs hierarchical agglomerative clustering using minimum variance method of Ward (Journal of American Statistical Association, 1963).

HICLUS -- This program is based upon an article in Psychometrika (1967) by S. C. Johnson. The program uses a single linkage and complete hierarchical agglomerative methods to analyze a similarity matrix which is input by the user.

HOWD -- This relatively simple iterative partitioning program uses a hierarchical diversive procedure to determine the number of clusters. Then the program iteratively uses reassignment passes until all entities are assigned to the cluster with the nearest centroid.

ISODATA -- ISODATA is based upon an analysis by G. H. Ball of techniques used in the pattern recognition literature which can be applied to clustering. ISODATA uses an iterative k-means procedure. The major options in the program control the splitting and/or merging of clusters. The final number of clusters is determined through a dynamic process.

MIKCA -- Like CLUS, MIKCA is concerned with iterative methods to optimize four statistical measures related to multivariate analysis of variance (McRae, 1971). The statistics are trace of W, determinant of W, largest eigenvalue of [(W**-1)*B] and the trace of [(W**-1)*B where W refers to the pooled within clusters covariance matrix and B is the pooled between-clusters covariance matrix. MIKCA allows the user to choose from three different measures of distance among entities.

III. Collections of Cluster Analysis Software Published in Books

ANDER -- Anderberg, M. R. Cluster Analysis for Applications. New York: Academic Press, 1973. Anderberg has written a general review of cluster analysis methods. In the appendices of the book are a number of FORTRAN subroutines which contain various clustering algorithms. There are two major types of methods covered in this book: hierarchical agglomerative methods (7 varieties) and iterative partitioning methods (3 varieties). The subroutines are relatively efficient to use in data analysis.

HART -- Hartigan, J. J. Clustering algorithms. New York: Wiley, 1975. Hartigan has written an introduction to cluster analysis methods which emphasizes a diversity of approaches to

this general family of statistical techniques. Major types of cluster analysis methods available in the book are hierarchical agglomerative, divisive, iterative partitioning, leader, and block clustering. Within each of these major types, the program options are relatively limited. Of the many discussions in the published scientific literature, Hartigan's book is the most innovative in its approach to cluster analysis.

The fourteen programs described above have been researched, and reports which describe the characteristics of these programs are available from the author. The reports are listed under the generic title "A Consumer Report on Cluster Analysis Software" with the subtitles

(1) Cluster Analysis Methods and their Literature
(2) Hierarchical Methods
(3) Iterative Partitioning Methods
(4) Useability

The reports can be obtained at cost from the author by writing to the following address:

Roger Blashfield
Department of Psychiatry
University of Florida
Box J-256, JHMHC
Gainesville, FL 32610

The final section of this paper is a subjective rank ordering of these programs in terms of their versability and useability. The rank ordering reflects the opinions and biases of this author. This rank ordering is based upon the versions of these program which were in existence as of 1 January 1977 and does not reflect updates of these programs after that time. Within each major subdivision of the ranking, the programs are listed in alphabetical order.

Excellent versatility, good useability
CLUSTAN

Good versatility, good useability
CLUS
NT-SYS

Good versatility, adequate useability
ANDER
BMDP
MIKCA

Either versatility or useability limited (but not both)
BCTRY

HART
HOWD

Minimal versatility and useability
BUILDUP
HGROUP
HICLUS
ISODATA
OSIRIS

References

Aldenderfer, M.S. & Blashfield, R.K. A survey of the use of cluster analysis in scientific research. Paper presented to the Classification Society, 23 May 1976.

Anderberg, M.R. Cluster analysis for applications. New York: Academic Press, 1973.

Bailey, K.D. Cluster analysis. In D. Heise (ed.) Sociological methodology. San Francisco: Jossey-Bass, 1974.

Bezdek, J.C. Cluster validity with fuzzy sets. Journal of Cybernetics, 1974, 3, 58-73.

Blashfield, R.K. Mixture model tests of cluster analysis: accuracy of four hierarchical agglomerative methods. Psychological Bulletin, 1976, 83, 377-388.

Blashfield, R.K. Questionnaire survey of cluster analysis software. Classification Society Bulletin, in press.

Cormack, R.M. A review of classification. Journal of the Royal Statistical Society (series A), 1971, 134, 321-367.

Cunningham, K.M. & Ogilive, J.C. Evaluation of hierarchical grouping techniques - a preliminary study. Computer Journal, 1972, 15, 209-215.

Everitt, B.S. Cluster analysis. London: Halstead Press, 1974.

Friedman, H.P. and Rubin, J. On some invariant criteria for grouping data. Journal of the American Statistical Association, 1967, 62, 1159-1178.

Hartigan, J. Clustering algorithms. New York: Wiley, 1975.

Hubert, L. Some extensions of Johnson's hierarchical clustering algorithms. Psychometrika, 1972, 37, 261-274.

Jardine, N. & Sibson, R. Mathematical taxonomy. New York: Wiley, 1971.

Johnson, S.C. Hierarchical clustering schemes. Psychometrika, 1967, 38, 241-254.

Ling, R.F. On the theory and construction of k-clusters. Computer Journal, 1972, 15, 326-332.

Lorr, M. & Radhakrishnan, B.K. A comparison of two methods of

cluster analysis. Educational and Psychological Measurement, 1967, 27, 47-53.

McRae, D.J. MIKCA: a cluster analysis program. Behavioral Science, 1971, 16, 423-424.

Peay, E.R. Nonmetric grouping: clusters and cliques. Psychometrika, 1975, 40, 297-313.

Sibson, R. Some observations on a paper by Lance and Williams. Computer Journal, 1971, 14, 156-157.

Sneath, P.H.A. and Sokal, R.R. Numerical Taxonomy. San Fransico: W.H. Freeman 1973.

Sokal, R.R. & Sneath, P.H.A. Principles of Numerical Taxonomy. San Francisco: W.H. Freeman, 1963.

Tryon, R.C. Cluster analysis. Ann Arbor: Edward Brothers, 1939.

Tryon, R.C. & Bailey, D.E. Cluster analysis. New York: McGraw-Hill, 1970.

Williams, W.T., Lance, G.N., Dale, M.B. & Clifford, H.T. Controversy concerning the criteria for taxonometric strategies. Computer Journal, 1971, 14, 162-165.

Wolfe, J.H. Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research, 1970, 5, 329-350.